Podcast episode transcript: Juggy Jagannathan and Thomas Schaaf

Juggy Jagannathan: Welcome to the Inside Angle Podcast from 3M Health Information Systems. This is your host, Juggy Jagannathan. My guest today is my friend and team lead, Thomas Schaaf. Thomas is a principal research scientist at 3M. He's a brilliant, super hardworking researcher working on the leading edge of speech and natural language understanding technologies. He received his doctorate from University of Karlsruhe. He has worked in LU Research labs in Sony, Toshiba, Amazon, and Carnegie Mellon University where he is currently an adjunct faculty at Language Technology Institute. Welcome to the podcast, Thomas.

Thomas Schaaf: Thank you, Juggy. Pleasure to be here.

Juggy Jagannathan: So, I was looking at your background. It's truly impressive. When did you get interested in this whole language technologies? Was your PhD related to this?

Thomas Schaaf: So, I actually got started in the speech recognition environment when I was a student in at the University of Karlsruhe, and I actually had to support myself. And there was this new professor, Alexander Waibel, who did the data collection and needed students to actually do the data collection. And so, I got a job even before I was okay to get a job there, and I found it super interesting.

Juggy Jagannathan: So yeah, I mean this is way back in early 2000 that you started working on speech, then?

Thomas Schaaf: That was in 1992.

Juggy Jagannathan: I didn't realize it got started that early.

Thomas Schaaf: Well, that was the data collection part, right? So, convincing students to read sentences for conference registration, but it was truly amazing to be part of just the data collection. And from that it actually evolved into more.

Juggy Jagannathan: Oh, you've been at it almost from the beginning.

Thomas Schaaf: I think it starts a bit earlier, but-

Juggy Jagannathan: It did. But the hearsay stuff and things like that. But still, 1990 is really early. I mean, what is it, 30 years ago?

Thomas Schaaf: It is. Yeah. Basically, I was working on what they now call the oil, right? Data is the new oil. At that time, data was very sparse, and it was also not easy to motivate students to actually donate their speech for this research. And the speech recognition at that time was also still quite in its infancy.

Juggy Jagannathan: It was like discrete word recognition, independent speech recognition. I mean there are all different versions of it in the early stages, I guess.

Thomas Schaaf: Yeah. I remember the very first, as a student project, way of doing, actually. Speech recognition was in a project that a colleague of mine, Monica Bohina was leading, and she made us implement what's called dynamic time warping. And that was already pretty amazing that you could actually do simple speech recognition with such a simple algorithm. And later, when I actually got deeper into this, attending classes by Alexander Bible. So, he had an on-hands lab, and that was awesome. So, we got to implement a lot of these algorithms from the ground up, which was also a great learning experience from a software engineering perspective. And that carried forward to another lab where we actually implemented a real speech recognition system again from the ground up. And that is actually how I found advisor for my master thesis to work on estimation of confidences for speech recognition outputs.

Juggy Jagannathan: So, you've been at it for a very long time. And your stints in Sony and Amazon, they all were related. Amazon, I guess you were working with the Alexa product?

Thomas Schaaf: That is correct. Yeah. So, I was in the Alexa team at the time. I think now you can say it, the Blueshift team. So, I was working on several speech recognition related products that all ended up with Alexa, the Fire TV, and some shopping devices that use speech recognition, and the mobile shopping app. So basically, I'm now in the former Amazon office, but there is now bigger Amazon office in Pittsburgh because I was the first and only employee at that time. So, at that time, I also were started working a little bit more in the natural language understanding side, or at least, the team that I was leading.

Juggy Jagannathan: Your efforts in Toshiba and Sony were also speech related?

Thomas Schaaf: So, in Toshiba, that was speech for speech-to-speech translation. So, English, Japanese to put that on the phone. But quite honestly, I was not very long at Toshiba because Detlef gave me call, and my wife said yes.

Juggy Jagannathan: We are certainly glad about that. And you also had a stint at CMU in the Language Technology Institute.

Thomas Schaaf: That was right after Sony. The story with Sony is that was a corporate research lab, and Sony had some difficult times. Don't want to go too much into this because for me, it was a very positive experience, so we had a lot of good collaborations going on because everybody got laid off, and that really inspired a lot of people to collaborate. And in the end, actually, we had very good demonstrations where we saved several jobs, not too many unfortunately. But yeah. I'd just finished my PhD. I was working the last year of my PhD at Sony writing my PhD thesis, basically on the train. That was a question was I should go to CMU or if I should try to go to another place like IBM or if I would try to stay with Sony. And I think working at CMU was really a great opportunity. I worked in speech recognition, again, that is my origin, on modern standard Arabic and Chinese speech recognition for DAPA projects.

Juggy Jagannathan: I noticed a trend where you have worked at various stages with multiple languages, and I noticed that you had this hackathon, and actually, it was this year, and I looked at all the projects. They were all focused on different languages and...

Thomas Schaaf: That is right. The theme for the hackathon was to build a community in especially this area where it took place, Middle East and Africa. The hackathon was part of the Spoken Language Technology Workshop 2022, which happened in 2023 because of the World Cup. And it was a great event. The motivation for the hackathon was to create a community of researchers and bring them together and build something for underrepresented languages. That is the reason why there are so many different languages.

Juggy Jagannathan: I was totally impressed with the range of projects as well as the representation. We had all the big tech representation, Google, Apple, Meta, Microsoft. They were all involved in this project in some fashion.

Thomas Schaaf: Yeah. The committee was pretty awesome. I have to say it was a lot of work, and without such a great committee, it would not have worked. Without the support of also the SLT larger committee would also have been difficult. It was a very positive experience. I think I can name a few people like Paula Garcia who really worked very hard with me on getting all the information to the students, solving all the problems. Because this is the first time we did such a hackathon on such a scale.

Juggy Jagannathan: Where is Paula from?

Thomas Schaaf: She's from Johns Hopkins University. And there were more other people like Nicole Boninelli who works on speech brain. And Chichi Batonabo who is at Carnegie Mellon University and Harshita Didi, who actually was very helpful to us because she was the person who had actually participated in hackathons.

Juggy Jagannathan: It really impresses me. Yeah, Johns Hopkins looks like they have something called a what, speech language institute? So, they've been doing this for quite some time also, right?

Thomas Schaaf: Yes. And actually, Harshita, we met there. She was a participant of one of these workshops over the summer working on machine translation. They are a great opportunity to bring researchers together to work on really interesting and hard problems and open new doors. And I think this is one of the fantastic organizations that really pushes the boundaries. I think we have been together there, and I think you also found it very interesting.

Juggy Jagannathan: Super awesome and really working on bringing industry and academy together in a cauldron to allow the students to innovate and try new things. That's really cool. I mean that's how research should be done, I guess.

Thomas Schaaf: Yeah. And hopefully, we have a team there one day, too. Still working on it. I think we have a good setup in a few years.

Juggy Jagannathan: You've also been working on the Language Technology Institute on the CMU side and a lot of people are just speech researchers or people like me are purely natural language technology focus, just understanding text. You are bridging the gap between the two. What do you think about the problems that we are currently working on, like working with doctor-patient conversation, which bridges that. We need to work with understanding conversation as well as making sure that we really do with it.

Thomas Schaaf: From my past experience is that speech recognition did not always work as well as it does now. And a lot of people think that it's actually a solved problem, but it is not. We still have to be able to work with a noisy input on the natural language understanding side. When we have doctor-patient conversations, it's in complicated situations from an environment. It's complicated from the kind of conversation that is going on. And the output is also not what we want to create. It's not a word-by-word transcript. Right? We want to actually produce a clinical note. And I have to say, the recent developments make me very optimistic, and this is exactly my research vision, that at one point these models will all work together more seamlessly, and there's good research going on at the moment to combine speech recognition with diarization, with summarization, and so on. There are still a lot of challenges to overcome, but I think we are on a very good way to accomplish that.

Juggy Jagannathan: I recently saw an article where speech recognition usually gets all this ha, hum, hmm, hmm, those kinds of almost inaudible perceptible noises, and it can misinterpret them. How can we deal with those kinds of problems when you're actually trying to create a clinical note?

Thomas Schaaf: That is one of the challenging questions. I think it's not the biggest challenge at the moment, but obviously the back channels sometimes have no information except that, "Oh yeah. Please keep talking." Right? You have to ignore them, and sometimes they actually mean a confirmation basically saying, "Okay, yes. I agree with you." And in the medical setup, there's two totally different things and they kind of sound the same, so it's really all in the context. And right now, I think everybody is using techniques that are basically requiring that you understand the context. And the newer networks that are currently developed are very good in picking up this nuances, if you have enough data. Once you have enough data, I think you make significant less confusing errors on that end.

Juggy Jagannathan: Can you talk about some of the results that we are having, even working with speech transcripts, which are not totally accurate.

Thomas Schaaf: I don't want to talk about papers that are currently in process to getting reviewed. So, dealing with speech recognition output, I think is really important. We have actually developed quite some understanding on how, for example, the physicians behave. We had this paper on actually detecting embedded dictations, which was a bit of a surprise to learn how much dictation actually still happens. Our long-term goal, of course, is to have the documentation based only on the conversations so that the doctor can reduce the amount of dictations they have to do.

And to some extent, or to large extent, the work on the clinical documentation actually creates a lot of effort is in sections like the history of present illness, which is totally conversational

and less structured in the output. And that actually creates effort on the physician to do that or a scribe. That's one of the reasons why we initially focused on these sections in our publications where we had the two-stage summarization to deal with the problems of very long input sequences, which technology is advancing. It becomes less of an issue. It's really interesting how did modeling and everything is changing in the last couple of, I would say years slash months. It's really fast at the moment.

Juggy Jagannathan: That's a great segue. Everybody is talking about ChatGPT and large language models, and I know we are riding the wave. We are actively exploring these technologies. What are your thoughts there? To me, it seems incredible, the advances you just mentioned about the technology moving fast and certainly has.

Thomas Schaaf: I think it's indeed a really interesting development that if you create a model that only predicts the next word, you can actually compress so much knowledge into these waits and also retrieve some of this knowledge. My feeling right now is you see a lot of good demonstrations or examples that are really stunning, and I think that is a technology that will probably over the next couple of years improve steadily. My take right now is that we need a proper and deep assessment on what it actually can do right now. And this is really just for making sure that if you want to use it, that it actually is safe to use. I'm very proud of the development that we are doing. These models are also very large, so there is from a business perspective, sometimes also a trade-off between cost to run these models.

Juggy Jagannathan: Yeah. Is it deployable.

Thomas Schaaf: Yeah. And how customizable it is. So, if you need several months to do prompt engineering and then something changes and you have to redo this, then it might be a difficult technology. I think it will be stable and it will get better, but at the moment, I would say there is a need to have a thorough assessment on what they can actually do. That also means that we have to look into this more deeply. Right?

Juggy Jagannathan: Yeah. And that is also a good segue. We just announced a collaboration with AWS on this front, and Amazon has entered this fray just as all the big tech has entered the fray of deploying these kinds of models. What are your thoughts there?

Thomas Schaaf: I think this is a really interesting development. It's a great opportunity, I think, to tap on AWS's experience. And I think it's also a great opportunity for them to work with us to see how a clinical note generation from doctor patient conversations can be done. It's in the early stages still, the collaboration. I think there's a lot of potential that this will help us to create better models, to create something that is actually faster, usable. They are working also on large language models. So, there are a lot of opportunities to assess and collaborate, I think, on model creation, model evaluation. We have discussion of joint experiments, which are still discussions. We have to figure out what is possible.

Juggy Jagannathan: Yeah. Of course, it's fairly in the early stages, and it's looks extremely promising. I just wanted to get back to this notion of guardrails. And this is something everybody who's using large language models are worried about, that it can produce stunning

results like you mentioned, but it can also go off rails. So, for us, the guardrails is using the scribe as the first mediator of our output. We are totally focused on the scribe at the moment in our deployment solutions. What do you think would best help the scribes?

Thomas Schaaf: Well, correct and complete notes in the way they want to see them. Right? I think speech recognition has to improve, and the note generation has to improve still. It's not done yet. So, we cannot go home.

Juggy Jagannathan: And less hallucinations. Or you don't like to use hallucinations. Less incorrect input.

Thomas Schaaf: The output of the notes has less factual errors and more relevant information. But overall, I think the models are quite good, and there are still a lot of problems that need to be solved, which I don't want to go into the details right now, but yeah.

Juggy Jagannathan: Oh, we have half a dozen papers in this area already published. And you are leading a super bright group of researchers. Any closing thoughts?

Thomas Schaaf: I think it's an exciting time. I have the feeling that the technology is evolving very fast, which I think sometimes can be quite daunting. But as you said, we have a great team of researchers that focus on making a contribution that really helps to improve the clinical notes for scribes. And then, long term, maybe we reach a quality that we can actually, for some notes can skip the scribe. Who knows? Right? So, that is still a pretty long way off, but I think this is possible. The future is bright.

Juggy Jagannathan: The future is bright. With that, thank you, Thomas.